

# Best Practices for Collecting, Managing and Curating Social Science Data: Final Report

Mitchell Davies  
Jeremy Kenyon  
Michelle Wiest  
J.D. Wulfhorst  
Marisa Guarinello

## Table of Contents

Executive Summary .....	3
Definitions.....	5
Introduction .....	6
Demographics .....	6
Researcher Needs, Behaviors, and Views.....	7
Current Approaches to Disclosure.....	14
Tools & Services .....	19
References .....	25
Appendices: Survey Methodology .....	28
Appendices: Institutional Review Board Approval .....	30
Appendices: NSF Social, Behavioral, Economic Sciences Data Sharing Policy.....	33

### Acknowledgements

We would like to acknowledge colleagues who made this project possible: Steven Daley-Laursen, Greg Gollberg, Snehalatha Gantla, and the University of Idaho Social Science Research Unit.

We would also like to recognize the organizations which helped provide funding for this work: the United States Geological Survey National Climate Change and Wildlife Science Center (USGS NCCWSC).

This work is licensed under a Creative Commons Attribution 4.0 International License.



## EXECUTIVE SUMMARY

The aim of this project and report is to facilitate expansion of current data management protocols to accommodate social science data for the USGS National Climate Change and Wildlife Science Center (NCCWSC) and its regional Climate Science Centers (CSCs). To address this expansion, we 1) identified the best practices and approaches from practitioners/experts through interviews with current curators of social science data, 2) explored the approaches of existing tools and services to determine if they are capable of meeting the needs of the NCCWSC, and 3) conducted a survey of the specific user community, with a focus on social science researchers funded by the NCCWSC and managers of the data within the program.

### Summary Conclusions

- While a majority of researchers indicate a belief in the importance of sharing data, many still do not. Sharing data is not yet universal nor the agreed experience of all in the social science research community.
- A majority of researchers believe data documentation is an important part of the research process, but few are interested in providing documentation past the content provided in a standard journal article.
- Those who work in social science data archives highlight lack of documentation as the single most common problem in curating social data.
- All data archivists interviewed provide some mechanism to address issues of sensitivity. Without access control, a formal risk analysis, or some other methods, data are not made available to the public.

Based on our review and analyses of the data collected during this project, we present recommended practices for researchers collecting data within the context of a federal mandate to share data, and recommend practices for an organization that is storing, curating, and providing access to data with human subjects sensitivities.

### Recommended Practices for Collecting Data:

- Research Design & Assurances
  - The proposed study must be reviewed by the researcher's institutional review board (IRB) to evaluate if the study is human subjects research.
  - If the study is deemed to be human subjects research, the researcher must provide a detailed description of study design and measures taken to protect privacy and mitigate risk to subjects to their IRB and obtain approval for this plan.
- Managing Data during the Project
  - Researchers should use data entry tools, such as RedCap, to ensure data fidelity.
  - Researchers must store data in a way that fulfills the mitigation strategies articulated in the human subjects research application to their IRB. This often includes storing data on a secure computer or server with access control to read or write to the databases.

- Before Submitting Data to an Archive
  - Researchers are responsible for performing data cleaning and de-identification or removal of sensitive data.
  - Researchers must create or provide the appropriate documentation for their data.
  - Researchers must indicate if, how, and conditions under which their data can be shared.

#### Recommended Practices for Archiving/Curating Data:

- Ingestion Procedures
  - Establish formal review procedures to determine if a depositor has engaged in appropriate disclosure reduction steps, especially but not only, removal of identifiers.
  - Require any reviewers to undergo training in risk assessment procedures, to gain familiarity with social science data and statistical disclosure control methods.
  - Obtain copies of protocols under which data were collected to identify conflicts with research oversight groups. Researchers may not be fully aware of the commitments they have made under local oversight, e.g. the IRB, and funder oversight.
- Documentation
  - Establish expectations for the creation of appropriate social scientific-oriented metadata. Identify necessary roles and responsibilities in relevant data policies.
  - Recognize that while much unique social science information – sampling methods, variable definitions (i.e., information often captured in codebooks), survey instruments and interview guides – can be embedded in non-social science metadata, it is not ideal. Researchers should be encouraged to use more relevant metadata standards, such as the Data Documentation Initiative (DDI) standard for survey data, where appropriate and feasible.
- Access Control
  - Every social science data archive possesses some form of data restriction. If restriction is not possible in a federal archive, a system of developing alternative representations of the data, i.e., excerpts rather than transcripts, must be identified, otherwise, the data should not be federally archived.
  - Data enclaves should be reviewed as a possible form of access control. This would place data in a physical location that can be accessed, but only by traveling there. The data are not otherwise networked.
- Tools and Services
  - Given the lack of familiarity with data documentation, provide a social science metadata editor for generating documentation to capture the research design elements of the project, as well as other information about the study necessary for interpretation of the data and re-use.

## DEFINITIONS

archiving	placing or storing something in a storage and retrieval facility or service for social scientific data
collapsing categories	combining groups or categories of a variable in order to reduce their number (collapsing, 2005)
data processing	conversion of items of information into a form that permits storage, retrieval, and analysis
deletion/anonymizing	involves more than simply removing the names of the participants under examination. It also involves removing or substituting all of the elements (e.g., names, places, and addresses) that might lead to the identification of an individual or group under examination (Lindsay & Goldring, 2010)
metadata	data that provides descriptive information (content, context, quality, structure, and accessibility) about a data product and enables others to search for and use the data product
quantitative data	information that can be measured and written down in numerical form
qualitative data	unstructured information that is either textual, or arranged into non-numerical categories
recoding/anonymizing	the process of making changes to the values of a variable (Tien, 2004)
risk of disclosure	the potential for identification of individuals to be made with or without any deliberate attempt to identify a person or organization
sensitive information	information that is or should be protected against unwarranted disclosure
standard	a specification for components, machines, materials, or processes intended to achieve uniformity, efficiency, and a specified quality
top-coding	a top-code for a variable is an upper limit on all published values of that variable. Any value greater than this upper limit is replaced by the upper limit or is not published on the microdata file at all (Statistics Netherlands, Statistics Canada, Germany FSO, and University of Manchester, 2005)
tools	specific programs designed to aid researchers in data collection or data management, normally reserved for computer or internet programs
services	specific programs designed to aid researchers in data collection or data management, normally reserved for programs that have involve human interaction
statistical disclosure control	the set of methods to reduce the risk of disclosing information on individuals, businesses or other organizations. Such methods are only related to the dissemination step and are usually based on restricting the amount of or modifying the data released (Statistics Netherlands et al., 2005)

## INTRODUCTION

In order for the field of science to uphold principles of transparency, openness, and reproducibility, policies and practices that incentivize the open sharing of data will be needed (Nosek et al., 2015). A shift to a more open science paradigm will require changes on the parts of government agencies, academic institutions, funding agencies/organizations, journals, and the researchers themselves. For example, journals are now beginning to require that data be shared in order for manuscripts to be published in accordance with guidelines created by the Transparency and Openness Promotion (TOP) Committee (Nosek et al., 2015). Major biophysical and social science journals are among the signatories to the TOP guidelines, including *Science*, *American Journal of Political Science*, *Behavioral Science and Policy*, *Political Science Research and Methods*, *Research and Politics*, *Sociological Science*, and *Survey Research Methods*, among many others (for a full list see: Center for Open Science, 2015). At the institutional level, changes will come in the forms of new policies and, to be effective, will be accompanied by training, tools, and other resources to help researchers meet these new policies. However, on the fundamental level, researchers will need to adopt practices within their own workflow that support open science. In considering such a shift toward open science, current knowledge about the needs, behaviors, and attitudes of researchers can be informative.

The results reported here provide insights on addressing the challenges of data management, sharing, and archiving in an emerging era of public accountability for scientific integration and transparency. The results from this project are organized as follows: We first summarize the characteristics of the interview and survey subjects (see Appendix 1 for a complete methodological description). We then delineate the needs, behaviors and views of the managers and researchers we surveyed or interviewed. Then, we describe current approaches to addressing these needs and meeting federal requirements. Finally, we provide an overview of tools and services available.

## DEMOGRAPHICS

### *Interview Respondents*

We interviewed 21 experts in the data management and data archiving fields to give us a base understanding of issues associated with current data sharing patterns and barriers. We selected the interviewees from a sample frame created from employees of top data management and archiving services (e.g. data managers, archive managers, documentation specialists, etc). These service organizations were primarily identified through the membership of the International Association for Social Science Information Services and Technology (IASSIST). The responses of these interviews were used to inform the development of the survey questions.

### *Survey Respondents*

To create a brief profile of survey respondents, the instrument included several ‘demographic’ questions to characterize respondent attributes related to disciplinary category, position, funding sources, and whether they considered themselves a climate scientist.

In summary, a profile of respondents includes:

- *Discipline type.* A majority (34/47 or 0.72) of the respondents identified as social scientists; others identified as either biophysical (5/47 or 0.10) or “other” (7/47 or 0.14: humanities, social and ecological science, social ecological systems, or interdisciplinary), and one respondent chose not to answer this question.
- *Position type.* The large majority of respondents (42/47 or 0.89) identified as employed in academic positions, while most of the remaining respondents (4/47 or 0.09) indicated they were federal employees. One respondent chose not to answer this question.
- *Funding sources.* Respondents indicated the sources of funding. The sampling frame provided a more diverse funding portfolio than only the USGS-related programs: USGS (18/47 or 0.38) vs non-USGS (29/47 or 0.62). The non-USGS funding sources were NSF (25/47 or 0.53), USDA (23/47 or 0.49), state agencies (17/47 or 0.36), private sources (12/47 or 0.26), NOAA (11/47 or 0.23), DOI (9/47 or 0.19), DOE (6/47 or 0.13), ‘other’ (6/47 or 0.13), EPA (4/47 or 0.08), and NASA (2/47 or 0.04). Two respondents chose not to answer this question.
- *Climate science.* When queried whether they considered themselves a “climate scientist” on the survey, a minority of respondents (7/47 or 0.15) indicated “Yes”, while the majority (39/47 or 0.83) of respondents indicated “No”. One respondent chose not to answer this question.

## RESEARCHER NEEDS, BEHAVIORS, VIEWS

### *Views on the Importance of Sharing Data*

Forty respondents shared their view of the importance of sharing data from the perspective of their discipline. Of these respondents, 24/27 or 0.60 ranked sharing data as important to some degree (15/47 or 0.37 very important, 9/47 or 0.23 moderately important), 9/47 or 0.23 described sharing data as rather unimportant (5/47 or 0.13 slightly important, not important 4/47 or 0.10) and 7/47 or 0.17 were neutral. When asked to answer the same question as it related to sharing *their own* data, 29/47 or 0.71 ranked sharing as important (18/47 or 0.44 very important, 11/47 or 0.27, moderately important), 8/47 or 0.19 described sharing data as rather unimportant (7/10 or 0.17 slightly important, not important 1/47 or 0.02) and 3/47 or 0.07% were neutral.

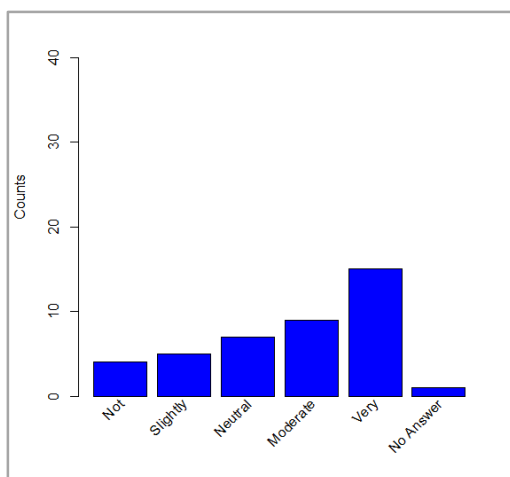


Figure 1: Importance of Sharing Data

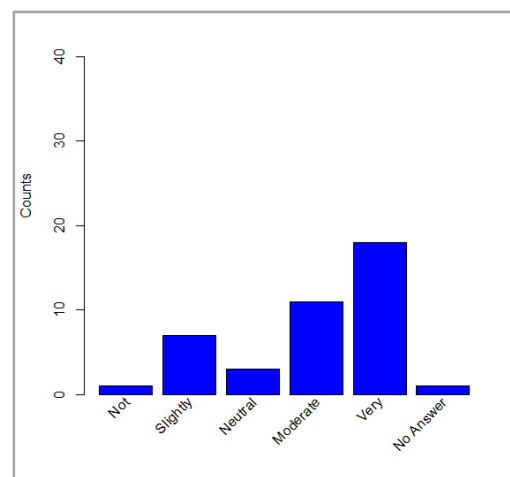


Figure 2: Importance of Sharing Your Data

Data documentation is important. Sharing data is not possible, or at least is far more difficult to accomplish, without proper documentation. Documentation ensures that those using data for secondary purposes will be able to learn and use important details to inform their own investigation. For example, details on data collection, cleaning, processing, and analysis are all critical to understanding the assumptions that underlie the data, in addition to basics such as units of measure, formulas, etc. A high percentage of the respondents were in agreement on the importance of documenting their data with 30/47 or 0.73 reporting that documentation of data as very important. Another 6/47 or 0.15 believe that data documentation is moderately important, 2/47 or 0.05 ranked it as slightly important and 2/47 or 0.05 were neutral on the subject. None of the respondents ranked data documentation as not important.

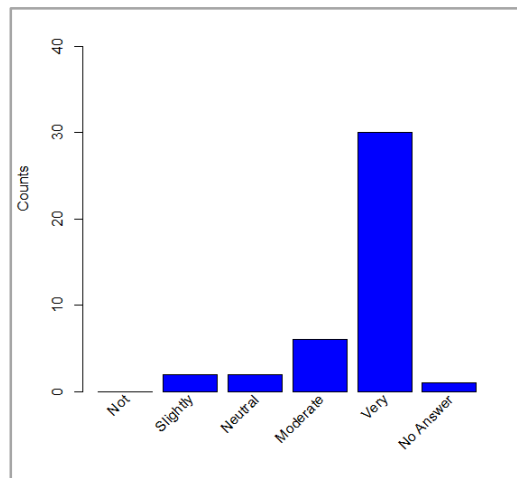


Figure 3: Importance of Documenting Your Data

### ***Researchers' Experience Sharing Data***

Data sharing is not universally practiced by researchers. In addition to researcher views on the importance of sharing and documenting data, we queried researchers about their experience sharing data. Of the respondents, 3/47 or 0.07 always share their data. Most researchers had moderate experience sharing their data, with 27/47 or 0.67 either sharing their data often (12/47 or 0.30) or sometimes (15/47 or 0.37). A quarter of the researchers have little to no experience sharing their data, 8/47 or 0.20 rarely share their data and 2/47 or 0.05 never share their data. The results for whether or not these researchers *use* shared data from others were fairly similar: 1/47 or 0.02 always, 10/47 or 0.24 often, 16/47 or 0.39 sometimes, 12/47 or 0.29 rarely, and 1/47 or 0.02 never use shared data.



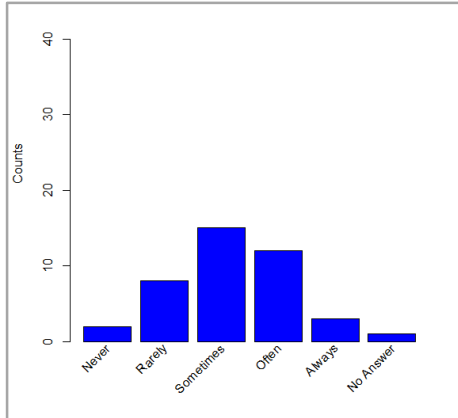


Figure 4: Experience Sharing with Others

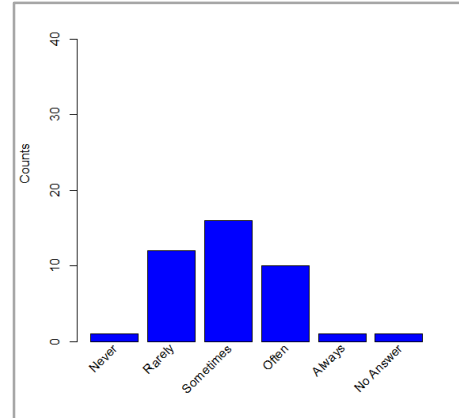


Figure 5: Experience Using Shared Data

One data manager we interviewed explained a key challenge associated with sharing data results from a lack of data curation experience in the social sciences:

*I think the biggest contribution we make to the research community is to require better curation from the researchers, but I don't think enough people do it. As far as I can tell, data curation and quality is not a top concern for most [social scientists]. We need to convince people that they should curate data.*

### ***Managing Qualitative Data***

Qualitative data deserves special recognition as its conventionally unstructured and intimate nature presents challenges different from numerical data. These data present preparation challenges for data sharing, especially when trying to maintain confidentiality and protect sensitive information. In fact, protecting confidentiality and privacy may require the removal not just of personally identifiable information (PII), but of contextual information required to understand the research setting (Broom, Cheshire, & Emmison, 2009). Removing this information is an effort to deal with “deductive disclosure”, the identification of an individual using known characteristics even though direct identifiers are removed (ICPSR, 2015). However, removing context introduces the question of the utility of data for secondary analysis and re-use. As Savage (2011) notes, “sources...abstracted from their original context... fail to adequately convey the subjectivities and identities of the research subjects...are neither standardized, nor are they adequately qualitative (p. 174)”. While these concerns are legitimate for re-users of qualitative data, they do not discount the efforts made by numerous organizations to archive and share this type of data. One data manager we interviewed explained that the research community varies in how it manages PII:

*...it can also be they just don't want to ever make it public if it's going to have many identifiers...they can remove them, but may consider it proprietary. Most people really want their data out there.*

Data managers also reported the research community may often not fully realize data management services and protocols in place:

*We have a guide to social science data preparation and we make that available but there is nothing required – just resources to help. We have policies, but the policies are on our end...We have a collection development policy, but that's nothing that 99% of anyone*

*would have read when making a (data) deposit. If you are the interested person, you can go and find it.*

In addition, data managers indicated that determining what the researchers find useful and accessible in de-identified datasets is challenging.

*...The most common issue that I run into is that researchers are not aware of what is in the data...We're very cautious about information we publicly release about the data. So, researchers have misconceptions about what is there from a technical perspective.*

In order to understand methods that we could use to address these challenges with qualitative data, we referred to the ICPSR's online guide for suggestions on possible ways to prepare qualitative data. The suggestions for preparing qualitative data only mentions confidentiality. There were two suggested methods for dealing with confidentiality and qualitative data: anonymizing the data, or deleting it. For anonymizing data, interviewees referred to taking out names and dates by either putting the person's relation to the subject, rather than the actual name (e.g., list 'uncle' instead of 'Uncle Bob'). The second suggestion was to delete only the sensitive part of the data, and use a reference mark to indicate that data had been deleted.

### ***Types of Data and Documentation Generated By Researchers***

Question 5, "What type of social science data are most commonly collected", was designed to differentiate between qualitative and quantitative social scientists, given the paradigmatic differences between the two approaches (Sandelwoski, 2009). Respondents (Figure 6) reported similar levels of qualitative (41/47 or 0.87) and quantitative (38/47 or 0.81) data collections. Based on the number of responses in each category, it appears that most researchers reported "mixed methods" approaches to their methodologies, or use of both quantitative and qualitative data. Not surprisingly, for environmental sciences-oriented research, geospatial data was also very common with 29/47 or 0.62 having reported collection of this data type. The other three categories – audio, video, and imagery – were reported as less often used, although some forms of imagery data could have gone in either the 'geospatial' or 'imagery' categories.

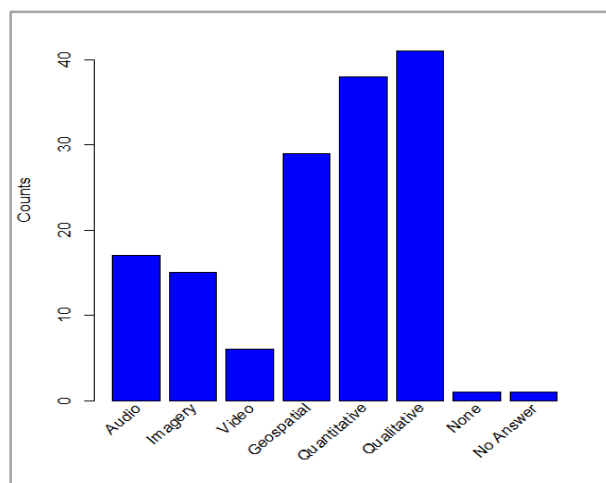


Figure 6: Types of data collected

A challenge of researchers reporting mixed methods in data collection is that the structure and mode of data collection can vary widely between, for example, general surveys and ethnographic observations. We sought clarification from respondents on what information is important to communicate to future users about the design of data collection through Q8: “When sharing social science research data, which design elements do you consider important to include?”

The responses (Figure 7) considered most important were related to structural elements of research design, including the method of collection (41/47 or 0.87), the location of where the data was collected (38/47 or 0.81), the selection criteria used to determine the sample (37/47 or 0.79), and the characteristics of the sample frame itself (30/47 or 0.64). Sample weights were not regarded as particularly important. Curiously, not a single respondent said that the mode of collection – phone, email, in-person – was important to communicate to future users.

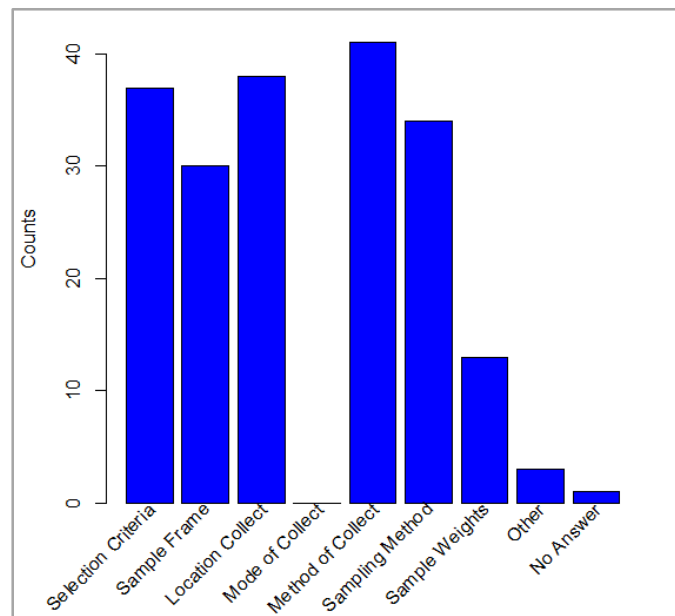


Figure 7: Importance of design elements

One interpretation of these results is that nearly all design elements are considered important, or moreover, that any design element may be important depending on the type of research being conducted. With this in mind, research design elements should undoubtedly be communicated to future users of a given dataset via documentation. With this in mind, we asked through question 9, “What type of documentation is considered the most important?” We sought clarification from researchers on what sorts of information is generally important for social science research data. The most frequent responses indicated the context of the study itself as the primary information of value. 38/47 or 0.81 of respondents replied that a “description of the study (summary, geography, date...)” was important to document. Respondents further reported that “research process information (data sources, methods, software/questionnaires...)” and the “people involved (authors, contributors...)” were also key at rates of 36/47 or 0.77 and 34/47 or 0.72, respectively.

By contrast, information about the dataset itself was considered less important. Data file descriptions, including format and data types, was considered important by a slim majority (24/47 or 0.51) of respondents. The codebook – used for explanations of variables, abbreviations, and similar elements of the data file – was only considered important by about a third of respondents (17/47 or 0.36). These

responses suggest a situation in which study context is considered valuable for creating documentation, but dataset explanation is not. Arguably, study context is already outlined to some extent in journal literature and project reports, yet often dataset explanation is not. Conventions in social sciences training have led to patterns of comfort levels with the familiar (e.g., study context), but inconsistency with areas that have not been traditional to the disciplines, such as sharing data.

Considering that metadata is a common form of preserving documentation with a dataset for future users, we asked in Q10: “What metadata standard is most commonly used?” Our goal was to verify the use of metadata standards, or the lack thereof. 29/47 or 0.62 of respondents said that they did not use any metadata (Figure 8). Considering that part of our sample contained USGS scientists, it was not surprising that a small number registered FGDC-CSDGM at 5/47 or 0.11. However, the second most frequent response was “no answer” at 6/47 or 0.13, which may be interpreted as unfamiliar with metadata schemas altogether, unable to answer the questions, or refusal to do so. The Data Documentation Initiative (DDI) schema, the only social science standard on our list, was reported as used by only 1 respondent.

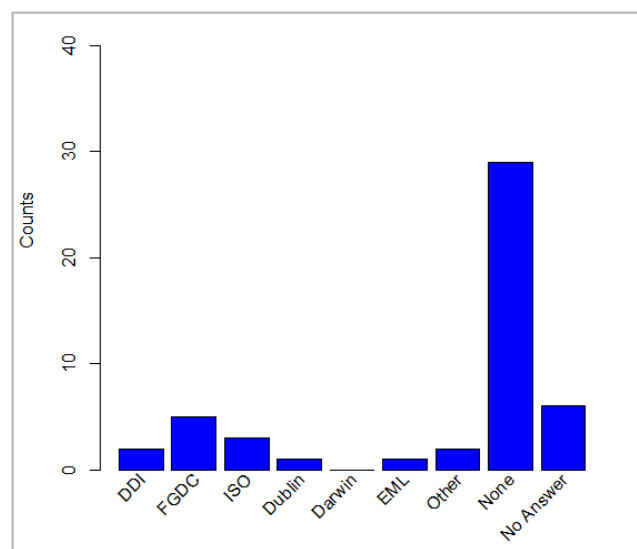


Figure 8: Metadata standards used

### ***Researcher’s Perceptions of Data Management***

While the previous questions helped outline the behaviors and practices of our respondents, we also sought to better understand their perceptions of the work involved with managing and curating data. Through question 14, we asked a multi-part question related to views on the responsibility for different parts of the process. We asked, “Although responsibilities for data management may overlap, which of these parties do you feel has *primary* responsibility for the data management processes listed?” We gave the respondent four parties with which to divide responsibility – a researcher, a funding organization, a data center, and no answer. Further, we gave them four phases of workflow in data management: providing access to data, creating documentation, performing data quality control (QC), and policy making.

Respondents perceived researchers (e.g., themselves) as responsible for the majority of the work, with 63.8% claiming that data quality control was their responsibility, 27/47 or 0.57 assigning responsibility to themselves for providing access to data, and 24/47 or 0.51 stating that creating documentation was their responsibility (Fig. 10). Only policy making was not seen as a researcher role. When split into USGS-

funded respondents and others (Fig. 9), we saw a similar pattern in which all but policy making was viewed as a researcher's responsibility. USGS respondents did, however, see providing access to data as slightly less their responsibility than the other respondents.

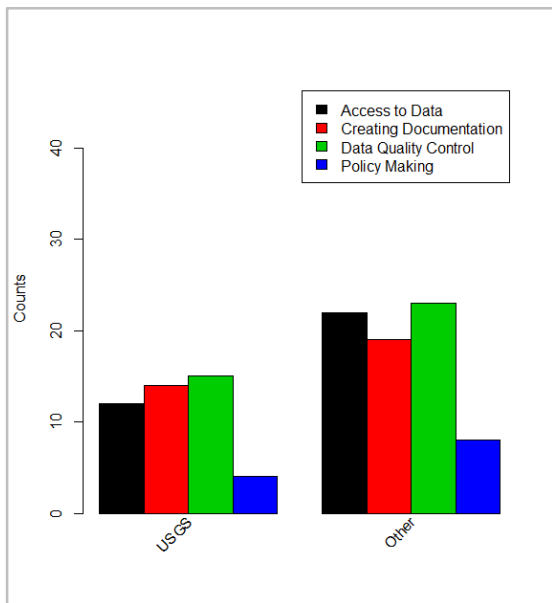


Figure 9: Role of Researcher: USGS funded vs. Other

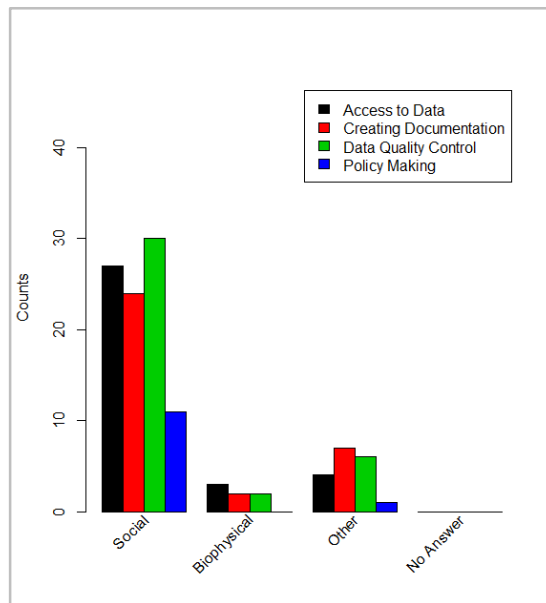


Figure 10: Role of Researcher: Social Scientists vs. Others

In terms of the roles of funding organizations (Figures 11 and 12), in neither breakdown did we see respondents point to a funding organization as having any responsibility other than policy making. In fact, for USGS-funded respondents, there was no role other than policy making for the funder.

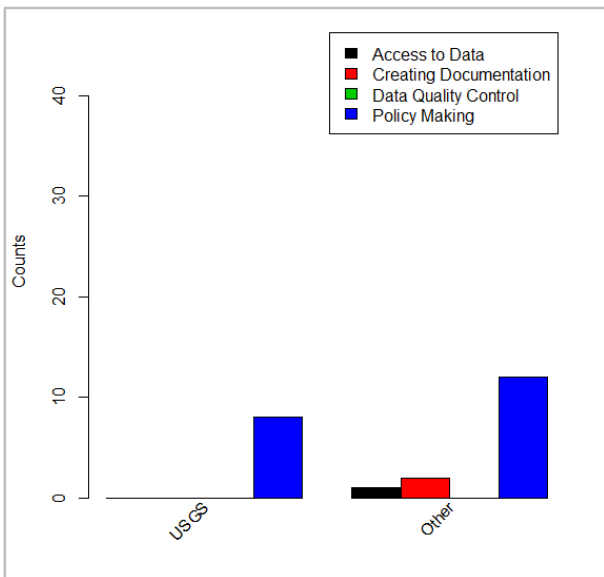


Figure 11: Role of funding agency: USGS funded vs. Other

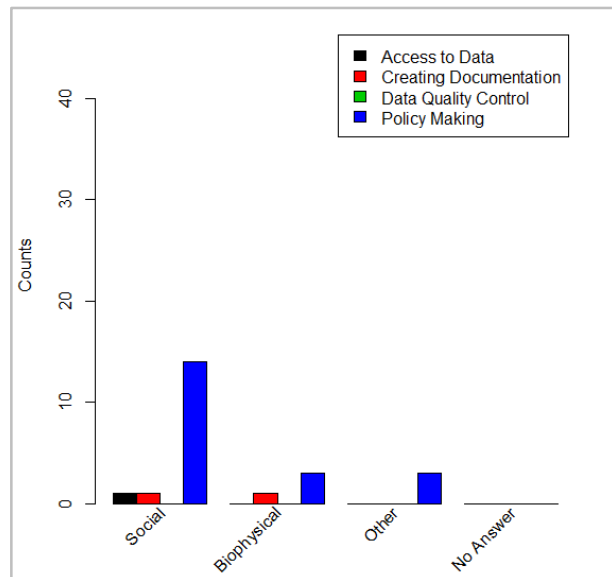


Figure 12: Role of funding agency: Social Scientists vs. Others

We found that data centers are likely underutilized and unrecognized as avenues for data sharing. Respondents indicated their perception of a small role for data centers. Like that of funding organizations, the role was perceived to be minimal at best (Figures 13 and 14). Related, data managers within our interviews described that researchers' lack of documentation can hinder the roles and services that data centers provide:

*We get mostly incomplete information – like it is missing the variables labels -- in addition to incomplete data....so basically, you are in the dark for creating metadata for the dataset....For instance, if there is poor annotation with a variable they created so that variable is not followed by the variable label and the value labels, then you don't know what the output would look like, and as the data manager, you would have a hard time following the flow of the research or the analysis.*

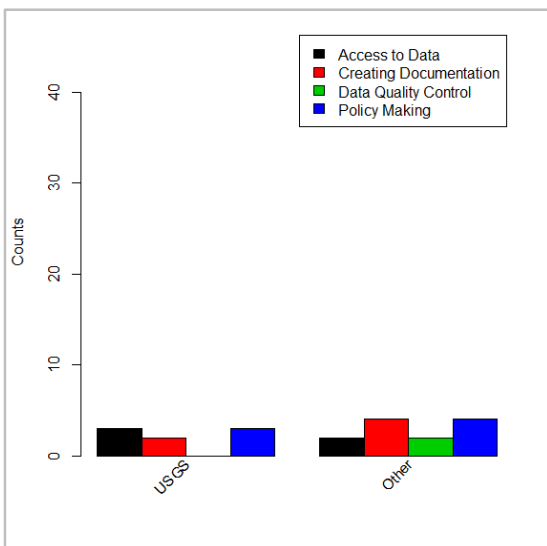


Figure 13: Role of data center: USGS funded vs. Other

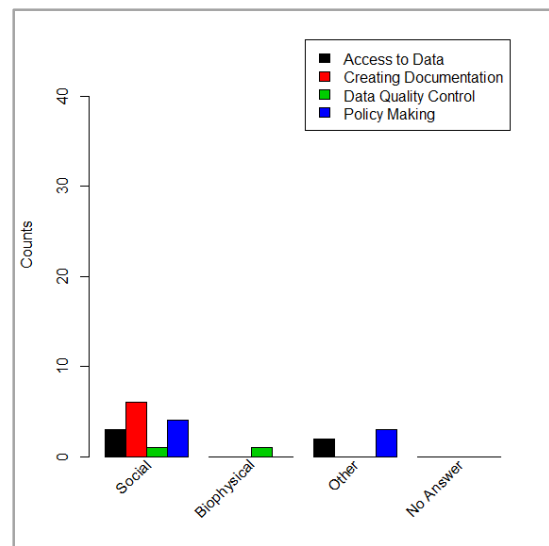


Figure 14: Role of data center: Social Scientists vs. Others

We infer that respondents perceive the work of data management lands primarily upon themselves due to a lack of familiarity with many resources available to them for data management and curation. This is evidenced by the lack of use and knowledge of metadata standards and lack of awareness of what services data center provide. Our interviews with data centers seem to suggest a strong focus on documentation and access mechanisms in support of social science. However, as shown with regards to qualitative research, re-use may demand the involvement of the original researcher, as qualitative data might be perceived as unusable without the original researcher's interpretative lens. In that light, access to the data, controlled by the originator, would depend on the researchers. That sort of local control would also then implicate local creation of documentation and local quality control.

## CURRENT APPROACHES TO DISCLOSURE

In this section, we identify a series of current approaches to meeting needs for managing data to prevent unintentional or unauthorized disclosure of sensitive information. We also include a review of current

standards of other federal agencies for regulation of collection and sharing of human-subjects data. This section is motivated by the White House mandate for federal agencies to increase public access to the results of research funded by the Federal Government. In a February 2013 memorandum, the Director of the Office of Science and Technology Policy required that all agencies granting over \$100 million in research expenditures develop a plan that includes a directorate to maximize data access and ensure that all researchers receiving federal grants and contracts develop data management plans (Holdren, 2013).

### ***Human Subjects Requirements and Data Sharing Policies of Other Federal Agencies***

#### *NIH and NSF*

The NIH and NSF define human subjects research as research involving a living individual about whom an investigator obtains either data through interaction or identifiable, private information (NIH, 2015b; NSF, no date). The policy under which both agencies work is what is referred to as the “common rule (NSF, no date).” The common rule is currently undergoing revision. Research is not considered human subjects research if it involves secondary analysis of coded data which was collected for some other reason and the investigator cannot readily ascertain the identity of the subjects (DHHS, 2014; NSF, no date).

Human subjects research is divided into two main categories: exempt or non-exempt from federal regulations. Types of human subjects research that are exempt include those studies conducted in educational settings, studies using non-identifiable, low-risk interviews or observations, or collection or study of existing data (NIH, 2015b). All else is considered non-exempt and require specific requirements be met in the application and during execution of the work. The NSF notes that social and behavioral scientists are subject to the same regulations as their biomedical colleagues (NIH, 2015b). In particular, investigators must address the following in their proposals:

- the involvement and characteristics of the study population, sources of materials, and potential risks,
- how the investigator will protect against risks through recruitment, consent, and additional protections for vulnerable populations,
- the potential benefits to human subjects and others, and weigh the risks in relation to benefits,
- the importance of the knowledge gained in relation to the risks (DHHS, 2014).

Both the NIH and NSF require grantee institutions to set up "Institutional Review Boards" (IRBs) to review research protocols and designs and ensure the protection of the rights of human subjects (NIH, 2015b; NSF, no date). Institutions determine whether the research is exempt or non-exempt and the extent of the IRB review required (NSF, no date).

To meet federal mandates, the NIH also collects race, ethnicity and sex count data on all funded human subject research using an inclusion management system (IMS) (NIH, 2001; NIH, 2015a). Investigators are required to submit a planned enrollment table prior to funding, and actual or cumulative enrollment through the life of the study (NIH, 2001). Grantees are able to access IMS through both the funding submission site (eRA commons) and the research performance progress report portal (NIH, 2015a). There, grantees enter information about their study, and fill out the enrollment table (NIH, 2015a).

The NIH also encourages data sharing, and states that all data should be considered for data sharing while safeguarding the privacy and confidentiality (NIH, 2012). It is required that any proposal requesting \$500,000 or more of direct costs in any year must include a data sharing plan (NIH, 2012). The data

sharing plan is to include a schedule for sharing, format of dataset, documentation, analytical tools, and mode of sharing. Data documentation and proper documentation is required to ensure proper use (NIH, 2012). Elements that should be included are methodology and procedures used to collect the data, details about codes, definitions of variables, and variable field locations (NIH, 2012). The NIH allows funds to be requested for data sharing and archiving (NIH, 2012).

The NSF requires a data management plan for ALL full proposals (NSF, 2010). Similar to the NIH data sharing plan, the NSF data management plans are meant to articulate what the primary research data and metadata will be, how it will eventually be shared and under what conditions, and how the data will be managed and maintained (NSF, 2010). Investigators are also encouraged to consider the legal and ethical restrictions on access to non-aggregated data and the scientific norms on data (NSF, 2010). The NSF monitors data management through annual and final reports. Particularly, they require investigators to discuss the execution of the data management plan (NSF, 2010). For more details on the NSF data management plan requirements, see the NSF publication “Data Management for NSF SBE Directorate” in Appendix Z.

#### USDA / NIFA & NOAA

In fall 2014, USDA published a white-paper report articulating the agency’s approach to increase access to “digitally formatted scientific data” resulting from its funded research programs and projects (USDA, 2014). USDA expects to formalize its data sharing policy during 2015-16, and explicitly recognizes lag within the agency compared to some other federal efforts. At this stage, data sharing practices are expected on the following timeline: a) *learning and expansion* (2015-16); b) *mainstream implementation* (2017); and c) *sustainable adoption* (2018 and beyond) (USDA, 2014).

In spring 2015, within USDA, the National Institute of Food & Agriculture (NIFA) articulated additional expectations to begin piloting a requirement for data management plans (DMP) within its competitive research and integrated programs. As noted within the agency’s ‘key concepts’ for data management, “accessing and sharing of digital and non-digital data helps increase the scope of scientific discoveries” (NIFA, 2015). In addition, the agency recognizes variability in data storage and archiving related to research type, disciplinary customs, and availability of financial resources. The DMPs must include specification of data repository plans to the extent known as well as strategies and contingencies about potential data loss or damage. Currently, the statement of expectations for data sharing and public access remains very general, and in light of this project’s results, arguably vague:

*Describe your data access and sharing procedures during and after the grant. Provide any restrictions such as copyright, patent, appropriate credit, disclaimers, or conditions for use of the data by other parties (NIFA, 2015).*

Related, the National Oceanic and Atmospheric Administration (NOAA) also released data access and sharing policies updating expectations for intramural projects and extramural grantees (NOAA Research Council, 2015). Revisions are expected to NOAA’s existing policy within 2015 and expected implementation in 2016 for external grantees. The emerging NOAA policy appears to emphasize the requirements for free access to data and results as well as the need for long-term archival solutions.

Both USDA and NOAA have very general statements with respect to human subjects research but expect ‘good stewardship’ and responsible conduct of research (RCR). The agencies indicate the expectation for



adherence to laws and policies to protect human research subjects should be standard protocol (NOAA, 2011; USDA, 2015), however, they do not state details as to what those laws and policies include, referencing only the Singapore Statement (Singapore Statement, 2010).

### ***Survey and Interview Results Regarding Disclosure Practices***

#### ***Assessing Risk***

Respondents to our survey indicated the use of risk assessment in the form of third-party reviews of their methods, processes, and data. Respondents who worked with American Indian Tribes reported relying on tribal review processes as a mechanism for disclosure control. Others relied on informed consent forms, normally developed within Institutional Review Board protocols, and another respondent identified ‘permission from original source’ as another type of informed consent.

#### ***Mitigating Risk***

While agencies have imposed a variety of rules regarding the handling of sensitive data, through our survey, we sought to better understand the work currently done by researchers to reduce the risk of disclosing sensitive data. We asked in our survey question 21: “Which processes do you use to reduce the risk of disclosure for sensitive data?” Based on our interviews, we identified a series of techniques used to reduce risk, including:

- *Top coding*
- *Deleting sensitive information*
- *Recoding/Anonymizing*
- *Collapsing categories*
- *Statistical disclosure control*

Of these choices, 24/47 or 0.51 of respondents said that they used recoding procedures to reduce risk of disclosure in their data. Following recoding, deletion of information was the next most commonly used at 19/47 or 0.40. 14/47 or 0.30 of respondents reported collapsing categories, and a few more (5/47 or 0.11) reported using other means of statistical disclosure control (Figure 15).

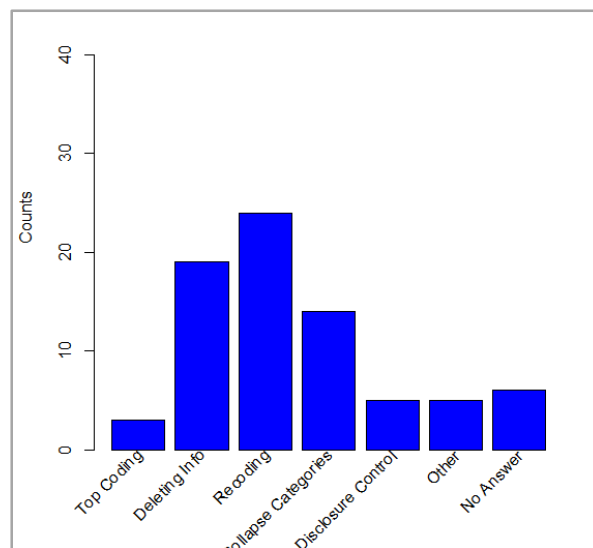


Figure 15: Methods of Disclosure Control Used

Some of the survey respondents volunteered further information through the free response sections of our survey. One respondent reported that they removed identifying information from qualitative data and another asserted that one should not attach names to shared files. Three other respondents promoted risk reduction through limiting access to the data, as one suggested limiting access to files and another respondent rejected the sharing of interview or qualitative data, except with trustworthy colleagues.

Our interviews with managers of data archives indicated use of similar practices regarding the process of managing sensitive data. The most common methods were recoding data, deleting risks, hiding variables, anonymizing data, completing a disclosure review, creating a modified public use version and restricting the data.

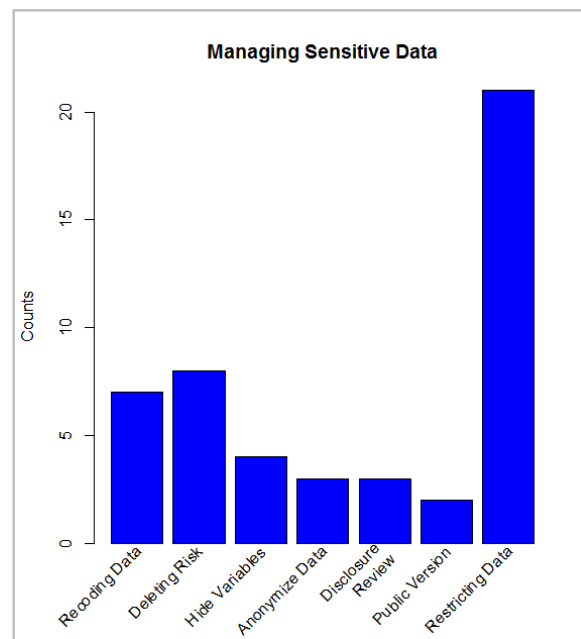


Figure 16: Methods of Handling Sensitive Data Among Archives

Given data restriction is a service provided by all interviewees, we sought further investigation of how they accomplish that role. Interviewees referred to a mediated process in which researchers requested permission for the data. Specifically 3/10 of those that responded this way noted that permission comes from the institution sharing the data, rather than the originator, and 2/10 referred to having password protection on the data. One-third indicated they put the project in offline storage, also referred to as a “data enclave, and one-third used the IRB as an approval mechanism.

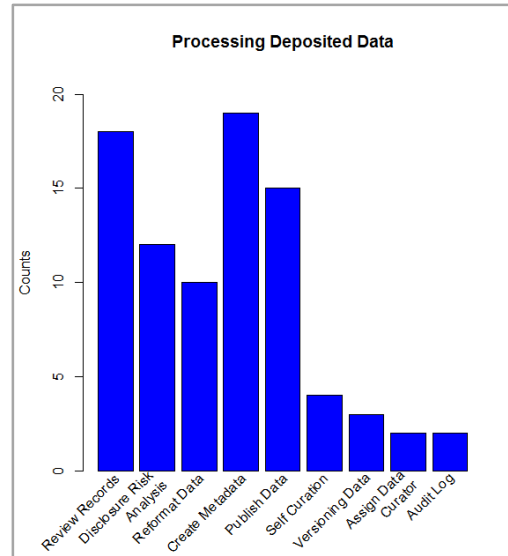


Figure 17: Procedures for Processing Deposited Data Among Archives

Interviewees also had clear procedures in many cases for processing deposited data and handling sensitivities. The most common procedures for sensitivity review were a review of the records (i.e., data, metadata, and codebooks), disclosure risk analysis on the deposited data, and reformatting the data. A smaller group creates multiple versions of the data.

For general archiving practice outside of risk assessment and sensitivity review, nearly all organizations create metadata records and engage in some sort of data publishing. Further, some assign data curators to the deposit process and maintain audit logs to track activity in processing the data. For those that conduct reviews of deposited data, specific review criteria include ensuring comprehensive data and metadata files, verification of data labels (performing cross-checks with the codebook, verifying that the data runs on various statistical packages, conducting disclosure reviews, and confirming that metadata is present. For treating problems in these reviews, respondents contact the data PI if problems with the data are found and reformat the data for access through SPSS, Stata, SAS, and other commonly used statistical programs.

## TOOLS & SERVICES

Given that social science data collection, management, and archiving is not just a contemporary concept and trend, we sought to better understand the available tools and services already existing and used by other organizations and researchers. Table 1 provides an overview of the strengths, weaknesses, and opportunities of numerous tools. Three dominant themes emerged:

- The first is that most tools contain a *mechanism for access control*. Access control features allow for multiple researchers to work on preparing the data together and making sure that each is only able to access the documents that they need.
- The second theme is the *creation of metadata documents/templates* which help researchers write and include metadata along with the data that is being archived. Features that assist with the latter

include allowing the user to share created metadata templates with others and options for cross-walking the metadata between standards.

- The third theme is the *uploading of documents* - in most cases, this is the mechanism by which data is being added into an archive.

Table 1. Tools for Social Science Data Management and Processing

Tools	Description	Strengths	Weaknesses	Opportunities
Colectica For Excel	Colectica for Excel is a free add-in for Microsoft Excel.	<ul style="list-style-type: none"> <li>• DDI-compliant metadata editor</li> <li>• Up-to-date</li> <li>• Easy side tab for documenting variables (i.e. columns in a table)</li> <li>• Can output DDI XML or save as a PDF or Word document</li> </ul>	<ul style="list-style-type: none"> <li>• Features are restricted, such as SPSS and Stata imports without purchasing the professional edition</li> <li>• Not much explanation or support to understand the different metadata fields</li> </ul>	<ul style="list-style-type: none"> <li>• Convenient way to introduce metadata creation into a researcher's working environment</li> </ul>
Colectica Repository	Colectica Repository is a centralized storage system for managing data resources, enabling collaborative workflows, and providing automatic version control.  <a href="http://www.colectica.com">http://www.colectica.com</a>	<ul style="list-style-type: none"> <li>• Contains DDI-compliant metadata editor</li> <li>• Software suite which covers research design through to deposition.</li> </ul>	<ul style="list-style-type: none"> <li>• Commercial Product</li> <li>• Training is available upon request</li> <li>• No user guides or tutorials without purchases</li> </ul>	<ul style="list-style-type: none"> <li>• Have many different programs which depend on the focus of the group you will have use the tool</li> </ul>
Dataverse	An open source web application to share, preserve, cite, explore and analyze research data.  <a href="http://dataverse.org/">http://dataverse.org/</a>	<ul style="list-style-type: none"> <li>• Stores metadata in DDI standard and can be cross walked into other metadata standards</li> <li>• Researchers can restrict specific files to only be downloaded by users with certain roles</li> <li>• Can upload any type of file with some types having more functionality</li> <li>• User permissions can be set to either change the permissions individually or as a group depending on how you set it up</li> </ul>	<ul style="list-style-type: none"> <li>• Very difficult to separate the different pieces of the application architecture. To use the metadata editor, you also must host the repository, or store it elsewhere.</li> </ul>	<ul style="list-style-type: none"> <li>• Organizations can host their own or use hosted versions of the application Able to create metadata templates and set these templates as default</li> </ul>
DDI Editor-Lite	DDI 3.0 Editor-Lite is an authoring tool created at ICPSR to support the production of DDI 3.0 Instances. It generates DDI 3.0-XML markup providing basic study and	<ul style="list-style-type: none"> <li>• Simple editor for generating DDI-compliant metadata</li> <li>• Web-based</li> <li>• No log in required</li> <li>• Can generate an XML file</li> </ul>	<ul style="list-style-type: none"> <li>• No save function</li> <li>• No storage of past files</li> <li>• No longer supported or maintained</li> </ul>	

	variable-level descriptions of simple, survey-type datasets.	<ul style="list-style-type: none"> <li>• Covers version 3.0</li> <li>• Validates field values</li> </ul>		
Nesstar	<p>Offers support for multilingual metadata, microdata, aggregate data, multi-layered maps, various visualization, subscriptions/notifications, cell notes/missing data symbols, basic analysis and embedding of live data into regular web pages</p> <p><a href="http://www.nesstar.com/">http://www.nesstar.com/</a></p>	<ul style="list-style-type: none"> <li>• Setups new users with specific roles easily</li> <li>• Access controlled by the roles set for the user</li> <li>• Contains multiple products – web viewer to work with data, publisher/editor for depositing data with metadata into a repository; repository software itself for management and curation</li> </ul>	<ul style="list-style-type: none"> <li>• Has specific ways to download different types of data</li> <li>• Is based on an older version of DDI for organization metadata, no clear updating scheme in place to get current</li> </ul>	<ul style="list-style-type: none"> <li>• Allows for creation of metadata templates and allow sharing metadata templates with other researchers</li> </ul>
Open ICPSR	<p>Premier social science data archive. Provides full support at all stages of the data curation process. Membership required.</p> <p><a href="https://www.openicpsr.org">https://www.openicpsr.org</a></p>	<ul style="list-style-type: none"> <li>• Easy to deposit data</li> <li>• Good user guides and staff support for depositing data</li> <li>• Researchers can restrict data depending on what is needed</li> </ul>	<ul style="list-style-type: none"> <li>• No control where data is published</li> <li>• \$600 deposit fee or membership (&lt;\$600) required</li> </ul>	
QDR	<p>The Qualitative Data Repository (QDR) is a dedicated archive for storing and sharing digital data (and accompanying documentation) generated or collected through qualitative and multi-method research in the social sciences.</p> <p><a href="https://qdr.syr.edu">https://qdr.syr.edu</a></p>	<ul style="list-style-type: none"> <li>• Easy to deposit data</li> <li>• Accepts many different types of files</li> <li>• Has a lot of options for restricting data</li> </ul>	<ul style="list-style-type: none"> <li>• They only take qualitative data in certain forms</li> </ul>	
RedCap	<p>A mature, secure web application for building and managing online surveys and databases</p> <p><a href="http://project-redcap.org/">http://project-redcap.org/</a></p>	<ul style="list-style-type: none"> <li>• Video tutorials as help documentation</li> <li>• Internal checks to make sure inputted data follows rules</li> <li>• Data validation rules can be set by researcher</li> <li>• Can set certain data to be identifiers and can create an identifier free version</li> <li>• Able to create survey and distribute it over email</li> <li>• Able to create longitudinal studies</li> </ul>	<ul style="list-style-type: none"> <li>• Does not create a metadata standard, just a codebook</li> <li>• Does not publish data just for data collection</li> <li>• Can be either hosted or you can set up your own instance, but either way requires technology/skill to set up</li> </ul>	

The tools and services above present a range of methods for managing, depositing and curating data. Some tools, like Dataverse, Nesstar, and the Colectica suite of products represent a holistic approach to serving researchers in a way that provides for documentation, access control, and risk management. These include upload tools, editing tools, tools to manage version control and access, and eventually, tools to enable access to the data based on customizable protocols. Others, such as OpenICPSR and QDR represent an organizational approach that may not be a solution for CSC-funded researchers, but represent a range of services an organization might provide. Depositing data via these organizations would represent a change in policy for USGS NCCWSC. They primarily present a researcher with a web form that are then sent to an internal data manager. From here, a series of interactions take place between the data manager and the researchers regarding the data.

Some tools did not classify cleanly into the above groups. Most of these address a specific problem with the data. For example, Redcap is primarily a data collection tool. The goal of using it is to provide researchers with a strong risk assessment and management environment, as to prevent disclosure of personal information during the collection process. Similarly, DDI Editor is primarily a metadata editing tool, albeit one that is no longer supported. It provides a graphic interface to enter data, and generates a DDI-compliant XML document on completion. It was most recently written to support the current DDI version (3.2), but may become out of date.

Overall, determining which tool to use depends on the level of involvement the researcher is willing to take in the preparation process. If the researcher is to take on the majority of the responsibility for preparing their own data for archiving, then tools which permit researcher to manage their own data would be most useful, such as RedCap, Nesstar Publisher, Colectica, or Dataverse. However, if the agency is willing to take on responsibility for review and assessment, then providing more narrowly focused tools, such as single metadata editors would be sensible.

### ***Survey and Interview Results Regarding Tools and Services***

To understand what researchers need in terms of tools we asked some questions in the survey referring to tools, specifically what formats researchers use, what tools researchers used, and what services researchers used. The formats are divided into 4 groups: 1) text formats which includes txt, pdf, docx and xml; 2) audio formats which includes wav and mp3; 3) picture formats which includes jpeg and tiff; and 4) statistical formats which include tabular, csv, R, SAS, tab, SPSS, STAT.

Table 2. Formats used by researchers for data files

Format types	Text formats				Audio formats		Picture formats	
	.txt	.pdf	.docx	.xml	.wav	.mp3	.jpeg	.tiff
%	36.2	55.3	80.9	27.7	29.8	44.7	72.3	36.2
Format types	Statistical formats							
	Tabular	.csv	R	SAS	Tab	SPSS	STATA	Other
%	59.6	51.1	23.4	21.3	2.1	46.8	27.7	8.5

These results indicate some formats are used more than others by the researchers we surveyed. For example docx was used by 38/47 or 0.81% of researchers, jpegs were used by 34/47 or 0.72 of researchers and the mp3 format was used by 21/47 or 0.45 of researchers. In contrast, there was not a single statistical format that was used more frequently by a large majority; however, there were three formats that indicated as the most frequently used: tabular/xlsx (28/47 or 0.60), csv (24/47 or 0.51) and SPSS (22/47 or 0.47).

Next, we look at the tools that are available and used by researchers. In order to know what type of tools and services are available to researchers, we asked data management experts in our interviews what tools and services they provide/recommend for researchers. Interviewees indicated they provide guidance for describing data management best practices, specific software, staff consultation services, data management plan creation guides/tools, and data storage and workspaces. Using this information we were able to compare different of tools with what researchers actually use for their research.

When we asked researchers about specific groups of tools that they use, we included the following choices as response categories: custom data management systems, documentation and metadata creation tools, archiving and publishing tools, data capture tools, online guides, public repositories, and secure web storage. Overall, the results indicate most respondents do not use most of these tools. Online storage (28/47 or 0.60), public repositories (20/47 or 0.43), and custom data management systems (CDMS) (17/47 or 0.36) were indicated as the most used tools (see Figure 18).

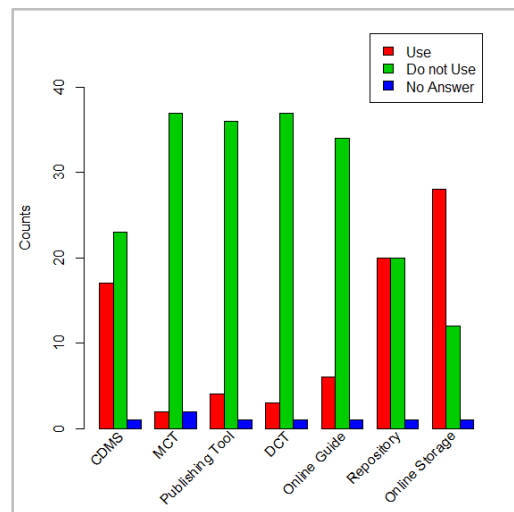


Figure 18: Most Commonly Used Categories of Tools Indicated by Survey Respondents

Finally, we asked researchers if they used the following services: staff consultation about data preparation, staff support for creating data management plans, training on data and metadata management, and webinar (Figure 19). As with the tools question, we found that more respondents do not use these services than those that do. However, the frequency of use of these services by researchers exceeds the rates of use for the tools we measured. There may be an advantage to providing human services to support researchers, and that building and providing tools may not be a solution in some cases.

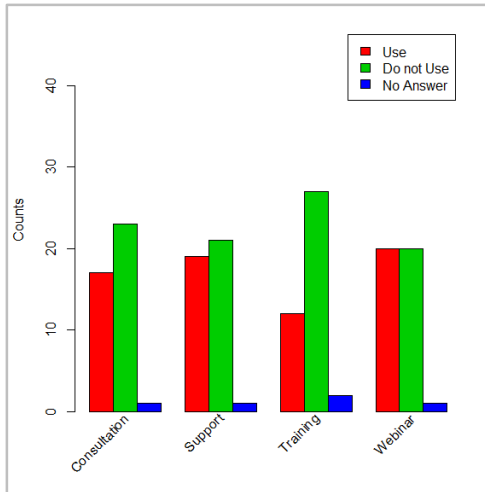


Figure 19: Most Commonly Used Categories of Services Indicated by Survey Respondents

## CONCLUSION

While a majority of researchers indicate a belief in the importance of sharing data, many still do not. Sharing data is not yet universal nor the agreed experience of all in the social science research community, and there is evidence of a lack of awareness of the resources available to facilitate sharing. A majority of researchers believe data documentation is an important part of the research process, but few are interested in providing documentation past the content provided in a standard journal article. Despite this, researchers and archivists are already aware and in agreement on appropriate approaches for risk mitigation and disclosure control, which is arguably the most important hurdle.



## References

- AAPOR (American Association for Public Opinion Research). (2015). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. Retrieved from: [http://www.aapor.org/AAPORKentico/AAPOR\\_Main/media/publications/Standard-Definitions2015\\_8theditionwithchanges\\_April2015\\_logo.pdf](http://www.aapor.org/AAPORKentico/AAPOR_Main/media/publications/Standard-Definitions2015_8theditionwithchanges_April2015_logo.pdf)
- Berg, B.L. (1995). *Qualitative Research Methods for the Social Sciences* (2<sup>nd</sup> ed.) Boston: Allyn & Bacon.
- Broom, A., Cheshire, L., & Emmison, M. (2009). Qualitative Researchers' Understandings of Their Practice and the Implications for Data Archiving and Sharing. In J. Goodwin (Ed.), *SAGE Secondary Data Analysis* (Vol. 43, pp. v4-173-v4-191). London: SAGE Publications Ltd. doi: 10.1177/0038038509345704
- Collapsing. 2005. In: W.P. Vogt (Ed.), *Dictionary of Statistics & Methodology* (3rd ed., p. 51). Thousand Oaks, CA: SAGE Publications, Inc. doi: <http://dx.doi.org/10.4135/9781412983907.n313>
- COS (Center for Open Science). (2015). *Transparency and Openness Promotion Guidelines*. Retrieved from: <https://cos.io/top/>
- DHHS (US Department of Health and Human Services). (2014, November). *Public Health Service Supplemental Grant Application Instructions for Competing Applications and Progress Reports*. Retrieved from <http://grants.nih.gov/grants/policy/hs/Preparing%20the%20Human%20Subjects%20Section.pdf>
- Holdren, J.P. (2013, February 22). *Increasing Access to the Results of Federally Funded Scientific Research*. Washington, DC: Executive Office of the President. Office of Science and Technology Policy. Retrieved from: [https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)
- ICPSR (Inter-university Consortium for Political and Social Research). (2015). *Deductive Disclosure Risk. Data Sharing for Demographic Research*. Retrieved from: <http://www.icpsr.umich.edu/icpsrweb/content/DSDR/disclosure.html>
- Lindsay, S., & Goldring, J. (2010). Anonymizing Data for Secondary Use. In A.J. Mills, G. Durepos, & E. Wiebe (Eds.), *Encyclopedia of Case Study Research* (pp. 25-27). Thousand Oaks, CA: SAGE Publications, Inc. doi: <http://dx.doi.org/10.4135/9781412957397.n10>
- NIFA (National Institute of Food & Agriculture). (2015, April). *Data Management Plan for NIFA-Funded Research Projects*. Retrieved from [http://nifa.usda.gov/sites/default/files/resource/Data\\_Management\\_Plan\\_NIFA\\_research\\_Apr\\_2015.PDF](http://nifa.usda.gov/sites/default/files/resource/Data_Management_Plan_NIFA_research_Apr_2015.PDF)
- NIH. (National Institutes of Health). (2001, August 8). *NIH Policy on Reporting Race and Ethnicity Data: Subjects in Clinical Research*. Retrieved from <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-01-053.html>

- NIH. (2012, February 9). *NIH Data Sharing Policy and Implementation Guidance*. Retrieved from [http://grants.nih.gov/grants/policy/data\\_sharing/data\\_sharing\\_guidance.htm#fin](http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#fin)
- NIH. (2015a, March 26). *Using the Inclusion Management System (IMS)*. Retrieved from [https://grants.nih.gov/grants/funding/women\\_min/inclusion\\_ims.htm](https://grants.nih.gov/grants/funding/women_min/inclusion_ims.htm)
- NIH. (2015b, July 23). *Research Involving Human Subjects*. Retrieved from <http://grants.nih.gov/grants/policy/hs/index.htm>
- NOAA (National Oceanic and Atmospheric Association) Research Council. (2011, December 7). *NOAA Administrative Order, 202-735D: Scientific Integrity*. NOAA Form 58-5 (4-04). Retrieved from [http://www.corporateservices.noaa.gov/ames/administrative\\_orders/chapter\\_202/202-735-D.pdf](http://www.corporateservices.noaa.gov/ames/administrative_orders/chapter_202/202-735-D.pdf)
- NOAA. (2015, February). *NOAA Plan for Increasing Public Access to Research Results*. Retrieved from [http://docs.lib.noaa.gov/noaa\\_documents/NOAA\\_Research\\_Council/NOAA\\_PARR\\_Plan\\_v5.04.pdf](http://docs.lib.noaa.gov/noaa_documents/NOAA_Research_Council/NOAA_PARR_Plan_v5.04.pdf) doi:10.7289/V5F47M2H
- Nosek B.A., Alter G., Banks G.C., Borsboom D., Bowman S.D., & Breckler S.J. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science* 348, 1422. doi: 10.1126/science.aab3847
- NSF (National Science Foundation). (2010, October 12). *Data Management for NSF SBE Directorate Proposals and Awards*. Retrieved from [http://www.nsf.gov/sbe/SBE\\_DataMgmtPlanPolicy.pdf](http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf)
- NSF. (no date). *45 CFR Part 690: Federal Policy for the Protection of Human Subjects. Subpart A: The Common Rule for the Protection of Human Subjects*. Retrieved from <http://www.nsf.gov/bfa/dias/policy/docs/45cfr690.pdf>
- Sandelowski, M., Voils C.I., & Knafl, G. (2015). On quantitizing. *Journal of Mixed Methods Research* 3(3), 208-222. doi:10.1177/1558689809334210
- Savage, M. (2011). Using archived qualitative data: Researching socio-cultural change. In J. Mason, & A. Dale (Eds.), *Understanding social research: Thinking creatively about method* (pp. 169-181). London: SAGE Publications. doi: 10.4135/9781446287972.n11
- Singapore Statement. (2010, July). *Singapore Statement on Research Integrity*. July, 2010. Retrieved from [http://www.singaporestatement.org/downloads/singapore%20statement\\_lettersize.pdf](http://www.singaporestatement.org/downloads/singapore%20statement_lettersize.pdf)
- Statistics Netherlands, Statistics Canada, Germany FSO, and University of Manchester. (2005, October 9-11). *Glossary of Statistical Disclosure Control*, incorporated in paper presented at Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva. Retrieved from: <https://stats.oecd.org/glossary/detail.asp?ID=7011>
- Tien, C. (2004). Recode. In: M. S. Lewis-Beck, A. Bryman, & Tim Futing Liao (Eds.), *The SAGE Encyclopedia of Social Science Research Methods* (pp. 931-932). Thousand Oaks, CA: Sage Publications, Inc. doi: <http://dx.doi.org/10.4135/9781412950589.n825>

USDA (U.S. Department of Agriculture). (2014, November 7). *Implementation Plan to Increase Public Access to Results of USDA-funded Scientific Research*. Retrieved from <http://www.usda.gov/documents/USDA-Public-Access-Implementation-Plan.pdf>

USDA. (2015). *Responsible and Ethical Conduct of Research*. Retrieved from <http://nifa.usda.gov/responsible-and-ethical-conduct-research>

### SURVEY METHODOLOGY

The Social Science Research Unit (SSRU) at the University of Idaho (UI) was subcontracted to conduct the web-based survey on best practices for data collection, archiving, sharing, and management among researchers. The sample was collected using snowball sampling procedures begun with NWCSC and NCCWSC funded researchers and other researchers known to be conducting related research on social aspects of climate science. During the survey, respondents were also prompted to extend the snowball sample with additional relevant colleagues also working in the specified area(s) of inquiry, but not yet in the sample frame. The final number of individuals recruited to the sample frame was 94. The UI Institutional Review Board reviewed and approved both the survey instrument and methodology for the project.

The study was divided into three different waves during spring 2015. The first respondent wave (N=69) were sent the email invitation during 8-10 April<sup>1</sup>. The second wave (N=11 respondents) received the email invitation on 21 April. The third wave (N=14 respondents) received the email invitation on 5 May. Within each wave, three rounds of subsequent reminder emails were sent to each non-responding individual at appropriate intervals following receipt of the survey. Data collection for all waves ended on 29 May. All survey data were collected using Sensus Web. Respondents completed the survey in an average of 15 minutes. Of those in the sample, 47 respondents completed the survey, including eight partial completes. The final response rate for the study overall was 50.0% percent. See Table 1 for a breakdown of dispositions and response rates by wave.

Table A-1. Final Survey Dispositions and Response Rates by Wave

Wave	Number of completes	Number of partial completes	Number of No-Action	Total	Response Rate <sup>2</sup>
1	30	7	32	69	53.6%
2	3	0	8	11	2.7%
3	6	1	7	14	50.0%
<b>TOTAL</b>	<b>39</b>	<b>8</b>	<b>47</b>	<b>94</b>	<b>50.0%</b>

### INTERVIEW METHODOLOGY

Initial primary data collection in the study began with a series of qualitative, in-depth interviews of experts in data management and data archiving. The objective of this phase focused on establishing a base of understanding trends and barriers in data sharing. A sample was collected using snowball sampling techniques (Berg, 1995) to generate an initial list of contacts (Directors, managers, or specialists) at leading organizations with data management and/or archive services. Supplemental suggested interviewees were added by those interviewed. Out of 43 individuals identified, interviews were conducted by telephone with 21 experts for this phase. Interviews lasted 30-40 minutes on average.

<sup>1</sup> Due to a technical error the first wave of respondents did not all receive the initial survey invitation simultaneously.

<sup>2</sup> Calculated using RR6 (p.53) from the American Association for Public Opinion Research (AAPOR, 2015).

The UI Institutional Review Board approved the interview guide (see below) and overall methodology for this project component (UI Protocol Code, #14-324; see below).

## University of Idaho Portal Navigation ▾

[Home](#) > [Protocols](#)

## Protocols

**About**

This page provides methods to input and manage research protocols in electronic format.

Start > Checklist > Setup > Exemption > Research > Data > **Review**

Please review the protocol below. Click Edit to change this record, click Delete to delete the record, or click Cancel / Close to return without saving changes.

Annotate

View: **Review Protocols** ▾

\* - indicates a required field

Cancel

Back

**Summary**

Please review the information you have submitted for this protocol.





Title:	<b>Best Practices and Approaches to Managing Social Science Data</b>
Status:	<b>Approved</b>
Protocol Type:	<b>IRB: Institutional Review Board</b>
Submission Type:	<b>New Application</b>
Project Type:	<b>Survey, Qualitative</b>
Other Type:	<b>N/A</b>
PI Expertise:	<b>PhD in field of research</b>
Other Expertise:	<b>N/A</b>
Purpose:	<b>N/A</b>
Design:	<b>N/A</b>
Procedures:	<b>N/A</b>
Research Subjects:	<b>N/A</b>
Privacy Level:	<b>Confidential means that the researcher will be able to link the subject's identity with his/her responses, but that this link will be maintained in a confidential manner.</b>
Exemption Categories:	<b>Category 2: Research involving the use of educational tests, survey procedures, interview procedures or observation of public behavior...</b>
Category Rationale:	<b>N/A</b>
Data Collection Methods:	<b>Self-Administered Survey, Personal Interview</b>
Other Data Collection Methods:	<b>N/A</b>
Experimental Method:	<b>N/A</b>

**Personnel**[Documents](#)[Confirmations](#)[Messages](#)[Amendments](#)[Renewals](#)[History](#)

This is a list of personnel associated with the selected protocol.

**PLEASE NOTE:** Applications **must** have a Primary Investigator (Faculty Sponsor) defined. Additionally, you **must** edit your own record below, and select your role.

Training certificates must be on file for all personnel (including yourself). Please note, training information may already be associated with some personnel, and do not need to be resubmitted.

	<input type="text" value="Quick Find"/>		 New Personnel		View: <span>Personnel</span>	
Showing 1-4 of 4 items   						
<input type="checkbox"/>	Username	Account Role	Account Type	Created On↓	Modified On	
	sneha	Investigator	UI Faculty / Staff	4/8/2015 10:33:53 AM	4/8/2015 10:34:07 AM	
	mwiest	Co-Primary Investigator	UI Faculty / Staff	7/8/2014 6:46:56 PM	7/8/2014 6:47:26 PM	
	davi6984	Student Investigator	Graduate Student	7/8/2014 6:42:48 PM	7/9/2014 8:22:49 AM	
	jd	Primary Investigator	UI Faculty / Staff	7/8/2014 6:42:26 PM	7/8/2014 7:01:10 PM	

## Data Management for NSF SBE Directorate Proposals and Awards

### Executive Summary

The National Science Foundation has released a new requirement for full proposal submissions regarding the management of data generated using NSF support. Starting in January, 2011, all proposals must include a data management plan (DMP).

The plan should be short, no more than two pages, and will be submitted as a supplementary document. The plan will thus not count toward the 15 page limit for proposals. The plan will need to address two main topics:

*What data are generated by your research?  
What is your plan for managing the data?*

“Data” are defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings. This includes original data, but also “metadata” (e.g. experimental protocols, code written for statistical analyses, etc.).

It is acknowledged that there are many variables governing what constitutes “data,” and the management of data, and each area of science has its own culture regarding data. The data management plan will be evaluated as part of your proposal. Proposals must include sufficient information that peer reviewers can assess both the data management plan and past performance. The plan should reflect best practices in your area of research, and should be appropriate to the data you generate. This document is meant to provide guidance for investigators within the Social, Behavioral, and Economic Sciences as they develop their Data Management Plans.

### Background

The National Science Foundation has released a new requirement for proposal submissions regarding the management of data generated using NSF support. Full proposals submitted, or due, to NSF on or after January 18, 2011 must include a data management plan (DMP). As summarized in the NSF Proposal and Award Policies and Procedures Guide:

*Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data ... created or gathered in the course of work under NSF grants.* ([http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag\\_index.jsp](http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_index.jsp), Section VI.D.4.b)

See the NSF [Grant Proposal Guide \(GPG\) Chapter II.C.2.j](#) for a description of the full policy implementation.

The full policy recognizes conditions under which restricting release of privileged or proprietary information would be appropriate, encourages sharing of software and inventions, and recognizes intellectual property rights. Dissemination of data is necessary for the community to stimulate new advances as quickly as possible and to allow prompt evaluation of the results.



## **The Requirement: Include a Data Management Plan in Proposals**

An appropriate data management plan is required as a supplementary document (maximum of two pages) for all full research proposals submitted. This plan is to be included in the Supplementary Documents section of the proposal and is not part of the 15-page limit for the Project Description. The NSF will not accept any full proposal submitted, or due, to NSF on or after January 18, 2011, that is lacking a DMP. Proposals submitted on or after January 18, 2011 for competitions with a target date prior to January 18, 2011 will require a DMP. Even if no data are to be produced, e.g. the research is purely theoretical or is in support of a workshop, a DMP is required. In this case, the DMP can simply state that no data will be produced.

The plan should describe how the PIs will manage and disseminate data generated by the project. The DMP will be considered by NSF and its reviewers during the proposal review process. Strategies and eventual compliance with the proposed DMP will be evaluated not only by proposal peer review but also through project monitoring by NSF program officers, by Committees of Visitors, and by the National Science Board.

NSF is aware of the need to provide flexibility in assessment of data management plans. In developing a plan, researchers may want to consult with university officials as many universities have explicit data management policies. Some professional organizations also have recommended data management practices (e.g. The American Economic Association at <http://www.aeaweb.org/aer/data.php>). A useful resource on preparing a data management plan can be found at ICPSR at <http://www.icpsr.umich.edu/icpsrweb/ICPSR/dmp/index.jsp>, including some very useful examples. Additionally, organizations that offer to store data may also focus on specific types of data. For instance, Open Context (<http://opencontext.org/>) and the Digital Archaeological Record (<http://www.tdar.org/>) provide data storage services for the archaeological community. NSF does not endorse the use of any specific repository.

## **Contents of the Data Management Plan**

The DMP should clearly articulate how “sharing of primary data” is to be implemented. It should outline the rights and obligations of all parties as to their roles and responsibilities in the management and retention of research data. It should also consider changes to roles and responsibilities that will occur should a principal investigator or co-PI leave the institution or project. Any costs should be explained in the Budget Justification pages. Specific components are listed below.

*Expected data.* The DMP should describe the types of data, samples, physical collections, software, curriculum materials, or other materials to be produced in the course of the project. It should then describe the expected types of data to be retained.

The Federal government defines ‘data’ in OMB Circular A-110 as:

Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This “recorded” material excludes physical objects (e.g., laboratory samples). Research data also do not include:

- |     |  |
|-----|--|
| (A) | Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and  |
| (B) | Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study. |

PIs should use the opportunity of the DMP to give thought to matters such as:

- The types of data that their project might generate and eventually share with others, and under what conditions
- How data are to be managed and maintained until they are shared with others
- Factors that might impinge on their ability to manage data, e.g. legal and ethical restrictions on access to non-aggregated data
- The lowest level of aggregated data that PIs might share with others in the scientific community, given that community's norms on data
- The mechanism for sharing data and/or making them accessible to others
- Other types of information that should be maintained and shared regarding data, e.g. the way it was generated, analytical and procedural information, and the metadata

*Period of data retention.* SBE is committed to timely and rapid data distribution. However, it recognizes that types of data can vary widely and that acceptable norms also vary by scientific discipline. It is strongly committed, however, to the underlying principle of timely access, and applicants should address how this will be met in their DMP statement.

*Data formats and dissemination.* The DMP should describe data formats, media, and dissemination approaches that will be used to make data and metadata available to others. Policies for public access and sharing should be described, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. Research centers and major partnerships with industry or other user communities must also address how data are to be shared and managed with partners, center members, and other major stakeholders.

*Data storage and preservation of access.* The DMP should describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. These can include third party facilities and repositories.

*Additional possible data management requirements.* More stringent data management requirements may be specified in particular NSF solicitations or result from local policies and best practices at the PI's home institution. Additional requirements will be specified in the program solicitation and award conditions. Principal Investigators to be supported by such programs must discuss how they will meet these additional requirements in their Data Management Plans.

## Post-Award Monitoring

After an award is made, data management will be monitored primarily through the normal Annual and Final Report process and through evaluation of subsequent proposals.

*Annual Reports.* Annual reports, required for all multi-year NSF awards, must provide information on the progress on data management and sharing of the research products. This information could include citations of relevant publications, conference proceedings, and descriptions of other types of data sharing and dissemination of results.

*Final Project Reports.* Final Project Reports are required for all NSF awards. The Final Project Report must discuss execution and any updating of the original DMP. This discussion should describe:

- Data produced during the award;
- Data to be retained after the award expires;
- Verification that data will be available for sharing;
- Discussion of community standards for data format;
- How data will be disseminated;
- The format that will be used to make data available to others, including any metadata; and
- The archival location of data.

*Subsequent proposals.* Data management must be reported in subsequent proposals by the PI and Co-PIs under "Results of prior NSF support."

## References and Resources

- Council on Governmental Relations, Access to and Retention of Research Data: Rights and Responsibilities, March 2006. <http://206.151.87.67/docs/CompleteDRBooklet.htm>
- National Science Foundation, Proposal and Award Policies and Procedures Guide, January 2010. [http://www.nsf.gov/publications/pub\\_summ.jsp?ods\\_key=papp](http://www.nsf.gov/publications/pub_summ.jsp?ods_key=papp)
- Office of Management and Budget, Circular A-110, September 30, 1999. White House Website, OMB Home. <http://www.whitehouse.gov/omb/circulars/a110/a110.html>